SYNTHETIC GENES FOR ENHANCED EXPRESSION

CROSS-REFERENCE TO RELATED APPLICATION

This application is a continuation-in-part and claims priority of Application No. 09/494,921; filed January 31, 2000.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

This invention was made in part with government support under Grant Nos. 1R43DK55951-01 and 1R43GM60822-01, awarded by the National Institutes of Health. The government thus has certain rights in the invention.

BACKGROUND OF THE INVENTION

The field of the invention is synthetic nucleic acid sequences for improved amplification and expression in a host organism, and methods of creating them.

It has been a goal of biotechnology to promote the expression of cloned genes for analysis of gene structure and function and also for commercial-scale synthesis of desirable gene products. DNA cloning methods have enabled the genetic modification of bacteria and unicellular eukaryotes to produce heterologous gene products. In principle, the genes may originate from almost any source, including other bacteria, animal cells or plant cells. Although this expression of heterologous genes is a function of a variety of complex factors, maximizing the expression of cloned sequences has been under intense and rapid development. Plasmid and viral vectors have been developed in both prokaryotes and eukaryotes that enhance the level of expression of cloned genes. In some cases the vector itself contains the regulatory elements controlling the expression of genes which are not normally expressed in the host cell so that a high level of expression of heterologous genes can be obtained.

Several problems exist, however, in the expression of many proteins across phyla and even across species. Post-translational handling and modification of expressed proteins by the host cell often does not mimic that of the heterologous gene's own cell type. Frequently, even if the protein is expressed in a useful form, heterologous genes are poorly expressed. Low yields of expressed protein may make manufacture of commercially useful quantities impossible or prohibitively costly. Vectors designed to enhance expression are not able to overcome some expression problems if the regulatory elements of the vector are not the constraint on robust expression. Other cellular or translational constraints are at issue.

Genes encoding poorly expressed proteins are often themselves difficult to clone and amplify as well. This can be due to secondary structure inherent in the gene, for example caused by high G-C content. Some methods have been used to reduce these difficulties, such as the use of DMSO or betaine to bring G-C and A-T melting behaviors more into alignment, or the use of ammonium sulfate (hydrogen binding cations) to destabilize G-C bonding during PCR. The problem with these methods is that the effects of the additives are concentration dependent, so variations in template size and G-C content mean lengthy optimization procedures. Additionally, these steps do nothing to facilitate subsequent expression of the nucleic acid once it has been cloned.

The frequency of particular codon usage in *Escherichia coli* and other enteric bacteria has long been known, and it has been hypothesized that replacement of certain rare codons encoding a particular amino acid in a heterologous eukaryotic or prokaryotic gene with a codon that is more commonly used by the selected host bacterium (or eukaryotic host cell) would enhance expression (see, e.g., Kane, *Curr Opin Biotechnol* **6:**494-500 (1995) and Zahn, *J Bacteriol.*, **178:**2926-2933 (1996)). This is based on the theory that rare codons have only a few tRNAs per cell and that transcription of heterologous sequences having numerous occurrences of these rare codons is limited by too few available tRNAs for those codons. However, simple replacement of rare codons does not reliably improve expression of heterologous genes, and no broadly applicable method exists to select which codon changes are best to increase expression of heterologous sequences. Further, it is not known in detail how codon usage is related to expression level.

Many gene products, often from bacteria, are commonly used as research and assay reagents, and various microbial enzymes increasingly are finding applications as industrial catalysts (see, for example, Rozzell, J.D., "Commercial Scale Biocatalysis: Myths and Realities," *Bioorganic and Medicinal Chemistry*, 7:2253-2261 (1999), herein incorporated by reference). Some have substantial commercial value. Examples include heat-stable *Taq* polymerase from *Thermus aquaticus,* restriction enzymes such as *Eco* RI from *E. coli*, lipase from *Pseudomonas cepacia,* β-amylase from *Bacillus sp.*, penicillin amidase from *E. coli* and *Bacillus sp.*, glucose isomerase from the genus *Streptomyces*, and dehalogenase from *Pseudomonas putida.* Genes from bacteria may express easily in commercially useful host strains, but many do not. In particular, genes from many bacteria have significantly different codon preferences from enteric bacteria. For example, filamentous bacteria such as streptomycetes and various strains of the genus *Bacillus, Pseudomonas,* and the like can be difficult to express abundantly in enteric

bacteria such as *E. coli*. An example of a *Pseudomonas* gene that is difficult to express in *E. coli* is the enzyme methionine gamma-lyase, useful for the assay of L-homocysteine and/or L-methionine as described in US Patent No. 5,885,767 (herein incorporated by reference). This assay is particularly useful in the diagnosis and treatment of homocystinuria, a serious genetic disorder characterized by an accumulation of elevated levels of L-homocysteine, L-methionine and metabolites of L-homocysteine in the blood and urine. Homocystinuria is more fully described in Mudd *et al.*, "Disorders of transsulfuration," In: Scriver *et al.*, eds., The Metabolic and Molecular Basis of Inherited Disease, McGraw-Hill Co., New York, 7th Edition, 1995, pp. 1279-1327 (herein incorporated by reference). In developing an assay for the accurate quantitation of L-homocysteine and L-methionine according to the methods described in Patent No. 5,885,767, obtaining large amounts of methionine gamma-lyase is necessary. However, this *Pseudomonas* gene contains a number of codons that are less commonly found in genes of desirable bacterial hosts for expression such as *E. coli*.

Similarly, genes from other organisms, such as yeast or mammals, can have utility as therapeutic agents, reagents, or catalysts. Examples include erythropoietin, human growth hormone, and eukaryotic oxidoreductases such as amino acid dehydrogenases, disulfide reductases, and alcohol dehydrogenases.

Because plasmid vectors designed to enhance expression with a variety of promotors or other regulatory elements often do not resolve the difficulty in expressing certain genes, and because no systematic approach exists for codon replacement to aid amplification of nucleic acids or their expression, there is clearly a need for an improved method for amplification and expression of genes, including genes from mammals and other animals, plants, yeast, fungi, and various bacteria such as streptomycetes, *Bacillus*, *Pseudomonas* and the like introduced into enteric bacterial hosts such as *E. coli*.

SUMMARY OF THE INVENTION

In one embodiment, the invention is directed to a method of making a synthetic nucleic acid sequence. The method comprises providing a starting nucleic acid sequence, which optionally encodes an amino acid sequence, and determining the predicted $\Delta G_{folding}$ of the sequence. The starting nucleic acid sequence can be a naturally occurring sequence or a non-naturally occurring sequence. The starting nucleic acid sequence is modified by replacing at least one codon from the starting nucleic acid sequence with a different corresponding codon to provide a modified nucleic acid sequence. As used herein, "codon" generally refers to a

nucleotide triplet which codes for an amino acid or translational signal (e.g., a stop codon), but can also mean a nucleotide triplet which does not encode an amino acid, as would be the case if the synthetic or modified nucleic acid sequence does not encode a protein (e.g., upstream regulatory elements, signaling sequences such as promotors, etc.). As used herein, a "different corresponding codon" refers to a codon which does not have the identical nucleotide sequence, but which encodes the identical amino acid. The predicted $\Delta G_{folding}$ of the modified nucleic acid sequence is determined and compared with the $\Delta G_{folding}$ of the starting nucleic acid sequence. In accordance with the invention, the predicted $\Delta G_{folding}$ of the starting nucleic acid sequence can be determined before or after the modified starting nucleic acid is provided.

Thereafter, it is determined whether the $\Delta G_{folding}$ of the modified nucleic acid sequence is increased relative to the $\Delta G_{folding}$ of the starting nucleic acid sequence by a desired amount, such as at least about 2%, at least about 10%, at least about 20%, at least about 30%, or at least about 40%. If the $\Delta G_{folding}$ of the modified nucleic acid sequence is not increased by the desired amount, the modified nucleic acid sequence is further modified by replacing at least one codon from the modified nucleic acid sequence with a different corresponding codon to provide a different modified nucleic acid sequence. These steps are repeated until the $\Delta G_{folding}$ of the modified nucleic acid sequence is increased by the desired amount to ultimately provide a final nucleic acid sequence, which is the desired nucleic acid sequence.

In one embodiment, the invention is a synthetic polynucleotide designed by the methods of the invention. This includes a nucleic acid having the sequence of a polynucleotide designed by the methods or a sequence complementary thereto.

In another embodiment, the invention includes a method of physically creating a tangible synthetic polynucleotide comprising creating a physical embodiment of the synthetic polynucleotide made using the nucleic acid/polynucleotide design methods of the invention, and the physical embodiments of the tangible synthetic polynucleotide prepared by this method (i.e., physical embodiments of the synthetic sequences, and copies of such sequences created by other methods.

The modified and/or final nucleic acid sequence can then be physically created. By the present invention, a desired nucleic acid sequence can be created that is more highly expressed in a selected host, such as *E. coli*, an insect cell, yeast, or a mammalian cell, than the starting sequence. By "more highly expressed" is meant more protein product is produced by the same host than would be with the starting sequence, preferably at least 5% more, more preferably at least 10% more, and most preferably at least 20% more.

Preferably the codon replacement is in a region of the starting nucleic acid sequence or modified nucleic acid sequence containing secondary structure. It is also preferred that the different corresponding codon is one that occurs with higher frequency in the selected host. In a particularly preferred embodiment, the desired amino acid sequence is expressed in *Escherichia coli*, and the amino acid sequence is from a bacterium of the genus *Pseudomonas*, and the different corresponding codon is selected to be one that occurs with higher frequency in a selected host, such as *Escherichia coli* than does the replaced codon. Alternatively, or in addition, the different corresponding codon is selected as one that has fewer guanine or cytosine residues than the replaced codon.

In a particularly preferred embodiment, the starting nucleic acid sequence is derived, e.g., converted, from an amino acid sequence native to an organism different from the desired host for expression, for example *Pseudomonas*.

The method of the invention also provides a modified, final sequence that is more amplifiable than the starting sequence. In other words, the final sequence is amplified more readily in a full length form, more rapidly or in greater quantity.

In another embodiment, the invention is directed to a synthetic nucleic acid sequence having a plurality of codons and encoding a methionine gamma-lyase protein from *Pseudomonas putida*. As used herein, the phrase "nucleic acid sequence encoding a protein" means that the nucleic acid sequence encodes at least the functional domain of the protein. The sequence having no more than about 95% homology, preferably no more than about 90% homology, more preferably no more than about 85% homology, still more preferably no more than about 80% homology, to a naturally occurring methionine gamma-lyase gene from *Pseudomonas putida*. At least about 5%, preferably at least about 10%, more preferably at least about 20%, still more preferably at least about 30%, even more preferably at least about 40%, of the codons in the synthetic nucleic acid sequence are different from codons found in the naturally occurring gene.

In one aspect, the codons in the synthetic nucleic acid sequence encode the same amino acids as the codons in the naturally occurring gene. In another aspect, at least one of the codons in the synthetic nucleic acid sequence encodes an amino acid different from the numerically corresponding amino acid found in the naturally occurring sequence. In yet another aspect, at least one of the different codons in the synthetic nucleic acid sequence is in an area of secondary structure in the naturally occurring gene.

In another embodiment, the invention is directed to synthetic genes derived from any source, e.g., eukaryotic or prokaryotic, for improved expression in heterologous or homologous expression hosts. The synthetic nucleic acid sequences of the invention are comprised on non-naturally occurring polymers of nucleic acids, each sequence having a biological function encoded by the sequence. The biological function can be direct (e.g., the nucleic acid sequence possesses the function, as in a promotor, for example) or indirect (e.g., the nucleic acid serves as a template to encode another molecule such as RNA or protein which has a function), and is generally one that is known from a similar naturally occurring or synthetic sequence. However, the biological function of the synthetic sequence created using the methods of the invention need not be identical to a known or predicted biological function in a known starting sequence. For example, the function may be enhanced in the synthetic sequence, or an enzyme may act on one or more different substrates, use more or different co-factors, catalyze reactions at a different rate, etc. The synthetic sequences further have no more than about 95% homology to a known starting sequence, and have a different free energy of folding than does the starting sequence. Finally, the synthetic sequences of the invention have the characteristic that they are better expressed (e.g., more highly expressed, expressed under different conditions, or expressed with more desired characteristics) in a selected host cell than the starting sequence would be if expressed in the selected host cell. The host cell is generally heterologous, but may be homologous for the starting sequence (the artificial synthetic sequence, not being found in nature, has no homologous host).

In one aspect, the synthetic nucleic acid sequence comprises a plurality of codons which encode amino acids and proteins. In preferred embodiments, the difference between the synthetic sequence and the starting sequence is that the synthetic sequence has at least one codon which is different from the starting sequence at the same amino acid position in the protein sequence. This codon may encode a different amino acid, the same amino acid, insert or delete an amino acid from that position, or encode a restriction site. Members of the oxidoreductase family are disclosed, and all members of this family or sequences encoding oxidoreductase functionality are among preferred sequences. Other preferred sequences include those encoding decarboxylase, formate dehydrogenase, hydantoinase, and vanillyl alcohol oxidase functions. Any sequence encoding a biological function from any source can be improved using the methods of the invention for enhanced expression or functionality.

In another embodiment, the invention is directed to a method of creating a synthetic nucleic acid. The method comprises providing a sense nucleic acid sequence having a 5' end and a 3' end and providing an antisense nucleic acid sequence having a 5' end and a 3' end. Preferably

the sense and antisense nucleic acid sequences are between about 10 and about 200 bases, more preferably between about 80 and about 120 bases. The 3' end of the sense sequence has a plurality of bases complimentary to a plurality of bases of the 3' end of the antisense sequence, thereby forming an area of overlap. Preferably the area of overlap is at least 6 bases, more preferably at least 10 bases, still more preferably at least 15 bases. The 5' end of the sense sequence extends beyond the 3' end of the antisense sequence, and the 5' end of the antisense sequence extends beyond the 3' end of the sense sequence. The method further comprises annealing the sense and antisense sequences at the area of overlap. A polymerase and free nucleotides are added to the sequences. Said nucleotides may be naturally occurring, i.e., A, T, C, G, or U, or they may be non-natural, e.g., iso-cytosine, iso-guanine, xanthine, and the like. The sequences can be annealed before or after addition of the polymerase and free nucleotides. The sequences are extended, wherein the area of overlap serves to prime the extension of the sense and antisense sequences in the 3' direction, forming a double stranded product. The extended sequence can then be amplified. Further, a second step to the method can be added where the double stranded first extension product is separated into an extended sense strand and an extended antisense strand and a second set of sense and antisense nucleic acid sequences are provided having a 5' end and a 3' end. Each has a plurality of bases on its 3' end complementary to a plurality of bases on the 3' end of the extended sense or antisense strand respectively, thereby forming second and third areas of overlap. A polymerase and free nucleotides are added to the sequences and separated strands, wherein the second and third areas of overlap serve to prime a second extension of the sequences and strands that encompasses the sequence of the first sense and antisense nucleic acid sequences and the second sense and antisense nucleic acid sequences.

DESCRIPTION OF THE DRAWINGS

These and other features and advantages of the present invention will be better understood by reference to the following detailed description when considered in conjunction with the accompanying figures wherein:

FIG. 1A: DNA sequence of synthetic *mdeA* gene (1200bps with GGT insertion), called *synmdeA*. *Nco* I and *Bam*H I cloning sites are engineered at 5' end and 3' end. The bold face uppercase nucleotides are the changed nucleotides from the original *mdeA* gene sequence.

FIG. 1B: First DNA segment, *mdeA1* (426 bps), with *Nco* I and *Pst* I cloning sites.

FIG. 1C: Second segment, *mdeA2* (414 bps), with *Pst* I and *Eco*R I cloning sites.

FIG. 1D: Third segment, *mdeA3* (367 bps), with *Eco*R I and *Bam*H I cloning sites.

FIG. 2A: First round of amplification using long oligonucleotides to generate template (*tpA1, tpA2,* or *tpA3*) DNA for each of the three *synmdeA* segments *mdeA1, mdeA2* or *mdeA3*. PCR amplification relies on overlapping sections of each oligonucleotide, which serves to prime the extension of the neighboring segment.

FIG. 2B: Second round of amplification using the two short oligonucleotides to amplify the full-length segments, *mdeA1, mdeA2* or *mdeA3*. The short oligonucleotides overlap with the 5' ends of the sense and antisense strands to form a template primed by the *tpA1, tpA2,* or *tpA3* strands, resulting in the filling in of both 5' and 3' ends of *mdeA1, mdeA2* and *mdeA3* after the second round of PCR.

FIG. 3 is a schematic of the cloning strategy for *mdeA1, mdeA2* and *mdeA3* into cloning and expression vectors. The amplified segments are ligated into the multiple cloning site of the illustrated vector in the top row, then *E. coli* are transformed with the plasmids. Individual plasmids containing each segment are selected in the second row, and the plasmids are double-digested to extract the insert, which is then ligated into an expression vector as shown in the last row.

FIG. 4A is a gel showing expression of a synthetic *P. putida* methionine gamma lyase *synmdeA* in BL21/pTM vector prior to and after induction with IPTG. All cultures were grown at 37°C. *synmdeA* was cloned into pET15b (available from Novagen) under the control of T7 RNA polymerase promotor. Lanes are: M - prestained protein molecular weight standards, high range, as indicated on the figure; 1 and 2 - three hours induction with 0.1 mM IPTG; 3 - three hours induction with 0.5 mM IPTG; 4 - three hours induction with 1 mM IPTG; 5 - three hours induction with 2 mM IPTG; 6 - not induced.

FIG. 4B is a gel showing the poor expression of native *P. putida* methionine gamma lyase (*mdeA*) in pSIT vector prior to and after induction with IPTG. All cultures were grown at 37°C. The induced samples contain extra bands at about 28 kD due to premature termination of *mdeA* translation. Native *mdeA* was cloned into the pSIT vector under the control of the T7 RNA polymerase promotor. Lanes are: M - prestained protein molecular weight standards, high range, as indicated on the figure; 1 - not induced; 2 and 3 - three hours induction with 0.5 mM IPTG; 4 and 5 - three hours induction with 1 mM IPTG.

FIG. 5 shows expression in *E. coli* of two genes with very different $\Delta G_{folding}$, naphthalene dioxygenase (NDO) from *Pseudomonas putida* (ΔG = -256.1 kcal/mol) and methionine gamma lyase (*mgl I*) from *T. vaginalis* (ΔG = -152.5 kcal/mol). Lanes 1-4 are NDO products, and 5-9 are MGL 1 products. Lanes are as follows: M1 - multimark multi-colored standard; M2 -

prestained protein molecular weight standards; 1 - not induced; 2- three hours induction with 0.02% L-arabinose; 3 - three hours induction with 0.04% L-arabinose; 4 - three hours induction with 0.08% L-arabinose; 5 - not induced; 6 - three hours induction with 0.02% L-arabinose; 7 - three hours induction with 0.04% L-arabinose; 8 - three hours induction with 0.08% L-arabinose; 9 - three hours induction with 0.10% L-arabinose. Both genes were cloned into the pBAD vector. Cells were grown at 37°C. Expression of *mgl I*, having a less negative $\Delta G_{folding}$ was superior to NDO expression.

FIG. 6 is a gel showing expression of native and synthetic genes developed using the methods of the invention. Lane 1 is a negative control (empty pBAD vector); Lanes 2 and 3 show expression of synthetic aldehyde reductase 2 containing an A25 to G25 mutation (synALR2mut) induced at 30°C and 37°C, respectively; Lanes 4 and 5 show expression of native yeast putative reductase 1 (YPR1) induced at 30°C and 37°C, respectively; Lanes 6 and 7 show the synthetic version, synYPR1, induced at 30°C and 37°C, respectively; and Lanes 8 and 9 show expression of synthetic aldehyde reductase 1 (synALR1) induced at 30°C and 37°C, respectively. All sequences except synALR1 were induced for 3 hours with L-arabinose. synALR1 was cloned into a different vector and induced for 3 hours with IPTG.

FIG. 7 is a gel comparing expression of native and synthetic formate dehydrogenase (Fdh1.2 and synFdh, respectively) induced with L-arabinose for 3 hours at 30°C and 37°C, and uninduced. Lane 1 is Fdh1.2 induced at 30°C; Lane 2 is synFdh at 30°C; Lane 3 is Fdh1.2 at 37°C; Lane 4 is synFdh at 37°C; and Lane 5 is uninduced Fdh1.2.

FIG. 8 graphically represents enzyme activity of synthetic formate dehydrogenase (synFdh) created using the methods of the invention (open triangles) as compared to native Fdh1.2 (open squares, induced; open circles uninduced), using an assay to catalyze the oxidation of formate in the presence of NAD^+.

DETAILED DESCRIPTION OF THE INVENTION

In one embodiment, the invention is directed to developing nucleic acid sequences that enhance expression of the encoded protein in a heterologous host. The frequency of particular codon usage for *E. coli* and other enteric bacteria is shown in Table 1, below. This table is derived from the 2000 Novagen Catalog, page 196, available online at http://www.novagen.com/html/catfram.html; herein incorporated by reference. However, the information in this table does not tell one of skill in molecular biology which codons should be replaced to enhance expression, if indeed any replacements will enhance expression.

Considerations other than simple codon replacement are clearly important. It has been discovered that the composition of the full gene (or fragment to be expressed) is more important than a particular codon exchange, and heterologous expression can be enhanced by replacement of codons in the sequence's open reading frame alone, independent of promotors or other regulatory sequence.

Table 1

aa	Codon	/1000[1]	Fraction[2]	aa	Codon	/1000[1]	Fraction[2]
Gly	GGG	1.89	0.02	Trp	UGG	7.98	1.00
Gly	GGA	0.44	0.00	stop	UGA	0.00	(stop)
Gly	GGU	52.99	0.59	Cys	UGU	3.19	0.49
Gly	GGC	34.55	0.38	Cys	UGC	3.34	0.51
Glu	GAG	15.68	0.22	stop	UAG	0.00	(stop)
Glu	GAA	57.20	0.78	stop	UAA	0.00	(stop)
Asp	GAU	21.63	0.33	Tyr	UAU	7.40	0.25
Asp	GAC	43.26	0.67	Tyr	UAC	22.79	0.75
Val	GUG	13.50	0.16	Leu	UUG	2.61	0.03
Val	GUA	21.20	0.26	Leu	UUA	1.74	0.02
Val	GUU	43.26	0.51	Phe	UUU	7.40	0.24
Val	GUC	5.52	0.07	Phe	UUC	24.10	0.76
Ala	GCG	23.37	0.26	Ser	UCG	2.03	0.04
Ala	GCA	25.12	0.28	Ser	UCA	1.02	0.02
Ala	GCU	30.78	0.35	Ser	UCU	17.42	0.34
Ala	GCC	9.00	0.10	Ser	UCC	19.02	0.37
Arg	AGG	0.15	0.00	Arg	CGG	0.15	0.00
Arg	AGA	0.00	0.00	Arg	CGA	0.29	0.01
Ser	AGU	1.31	0.03	Arg	CGU	42.10	0.74
Ser	AGC	10.31	0.20	Arg	CGC	13.94	0.25
Lys	AAG	16.11	0.26	Gln	CAG	33.83	0.86
Lys	AAA	46.46	0.74	Gln	CAA	5.37	0.14
Asn	AAU	2.76	0.06	His	CAU	2.61	0.17
Asn	AAC	39.78	0.94	His	CAC	12.34	0.83
Met	AUG	24.68	1.00	Leu	CUG	69.69	0.83
Ile	AUA	0.15	0.00	Leu	CUA	0.29	0.00
Ile	AUU	10.16	0.17	Leu	CUU	3.63	0.04
Ile	AUC	50.09	0.83	Leu	CUC	5.52	0.07
Thr	ACG	3.63	0.07	Pro	CCG	27.58	0.77
Thr	ACA	2.03	0.04	Pro	CCA	5.23	0.15

aa	Codon	/1000[1]	Fraction[2]	aa	Codon	/1000[1]	Fraction[2]
Thr	ACU	18.87	0.35	Pro	CCU	2.76	0.08
Thr	ACC	29.91	0.55	Pro	CCC	0.15	0.00

[1] Expected number of occurrences per 1000 codons in enteric bacterial genes whose codon usage is identical to that compiled in the frequency table.

[2] Fraction of occurrences of the codon in its synonymous codon family.

The present invention encompasses highly amplifiable, expressible oligonucleotides, polynucleotides, and/or genes and is directed to methods of designing and physically creating these nucleic acid sequences. In one embodiment, the present invention is directed to a method of designing and physically creating genes that express well when introduced into heterologous expression hosts, such as from eukaryotic sources into prokaryotic hosts, e.g., common enteric bacterial host microorganisms such as *E. coli*. The invention allows expression of genes from various organisms, such as mammals and other animals, plants, yeast, fungi, and bacteria (e.g., pigs, *Saccharomyces*, streptomycetes, *Bacillus, Pseudomonas* and the like) in prokaryotic hosts such as *E. coli* and eukaryotic hosts at commercially viable levels, even proteins with typically low yields, such as methionine gamma-lyase from *P. putida*. As used herein, the terms "polypeptide," "protein" and "amino acid sequence" are used interchangeably and mean oligomeric polyamides of at least two amino acids, whether or not they encompass the full-length polypeptide encoded by a gene or merely a portion of it. "Heterologous" indicates that the sequence is not native to the host used or identical to a sequence which naturally occurs in the host used, or refers to a host which is not the natural source of a nucleic acid or peptide sequence. "Designing" means conceiving a sequence of nucleotides in a form that can be written or printed. Such sequence may correspond to the coding region of an entire gene, or only a portion of it, and may also include additional bases added at a particular location or position, for example to create desired restriction sites or to insert mutations to enhance the protein's function. "Physically creating" means preparing a chemical entity such as an oligonucleotide/polynucleotide or polypeptide, whether by synthesis by chemical and/or enzymatic methods, biosynthesis, a combination of synthesis and PCR, or by any other methods known in the art. "PCR" means polymerase chain reaction.

In the present invention, the sequence of a gene is modified to enhance its ability to be amplified, for example by PCR methods, and/or to improve its expression in a selected host, for example, an enteric bacterium such as *E. coli*. This is achieved by designing a nucleotide sequence preferably using codons preferred by the host, calculating the $\Delta G_{folding}$ of the nucleic

acid sequence (the amount of energy required for or released by folding in solution, in kcal/mole), modifying the sequence by replacing one or more codons in the sequence in one or more areas of predicted secondary structure with less preferred codons to reduce predicted secondary structure, and recalculating the $\Delta G_{folding}$ of the modified nucleic acid sequence. The replacement of codons and recalculation of the free energy of folding may be repeated as many times as desired. One, some, or all codons encoding a particular amino acid may be replaced in the region of secondary structure, or throughout the entire coding region of the sequence. The result is a modified final nucleic acid sequence, for example a synthetic gene encoding a desired complete or partial protein, whether a mutant protein or one having the desired structural and functional attributes of a native protein. The final synthetic sequence may be optimized for only a single selected host, but the methods of the invention are readily operably for a starting sequence from any source for expression in any selected host, whether animal, plant, fungal, prokaryotic, etc.

As used herein, the term "synthetic" gene, nucleic acid, oligonucleotide, polynucleotide, primer, or the like means a nucleic acid sequence that is not found in nature; in other words, not merely a heterologous sequence to a particular organism, but one which is heterologous in the sense that it has been designed and/or created in a laboratory, and is altered in some way, and that it does not have exactly the nucleotide (or possibly amino acid) sequence that its naturally occurring source, template, or homolog has. A synthetic nucleic acid or amino acid sequence as used herein can refer to a theoretical sequence or a tangibly, physically created embodiment. It is intended that synthetic sequences designed by the method be included in the invention in any form, e.g., paper or computer readable ("theoretical"), and physically created nucleic acids or proteins. Physically created nucleic acids and proteins of the invention are part of the invention, whether derived directly from the designed sequence, or copies of such sequences (e.g., made by PCR, plasmid replication, chemical synthesis, and the like). The term "synthetic nucleic acid" can include, for example, nucleic acid sequences derived or designed from wholly artificial amino acid sequences, or nucleic acid sequences with single or multiple nucleotide changes as compared to the naturally occurring sequence, those created by random or directed mutagenesis, chemical synthesis, DNA shuffling methods, DNA reassembly methods, or by any means known to one of skill in the art (see e.g., techniques described in Sambrook and Russell, "Molecular Cloning; A Laboratory Manual," 3rd Ed., Cold Spring Harbor Laboratory Press (2001), herein incorporated by reference). Such alterations can be done without changing the amino acid sequence encoded by the nucleic acid sequence, or can modify the amino acid sequence to leave a desired function of the encoded protein unaltered or enhanced. As used herein, "nucleic acid" means a naturally

occurring or synthetic nucleic acid, which can be composed of natural or synthetic nitrogen bases, a deoxyribose or ribose sugar, and a phosphate group.

"Secondary structure" refers to regions of a nucleic acid sequence that, when single stranded, have a tendency to form double-stranded hairpin structures or loops. Such structures impede transcription (or amplification *in vitro*) and translation of affected regions in the nucleic acid sequence. Nucleic acids can be evaluated for their likely secondary structure by calculating the predicted $\Delta G_{folding}$ of each possible structure that could be formed in a particular strand of nucleic acid. Energy must be released overall to form a base-paired structure, and a structure's stability is determined by the amount of energy it releases. The more negative the $\Delta G_{folding}$ (i.e., the lower the free energy), the more stable that structure is and the more likely the formation of that double-stranded structure.

Computer programs exist that can predict the secondary structure of a nucleic acid by calculating its free energy of folding. One example is the *mfold* program, which can be found at http://mfold2.wustl.edu/~mfold/dna/form1.cgi (using free energies derived from SantaLucia *Proc. Natl. Acad. Sci. USA* **95:**1460-1465 (1998); see also Zuker, *Science*, 244, 48-52, (1989); Jaeger *et al.*, *Proc. Natl. Acad. Sci.* USA, Biochemistry, **86:**7706-7710 (1989); Jaeger *et al.*, Predicting Optimal and Suboptimal Secondary Structure for RNA. in "Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences", R. F. Doolittle ed., *Methods in Enzymology*, 183, 281-306 (1989); all herein incorporated by reference). Another example of such a computer program is the Vienna RNA Package, available at http://www.ks.uiuc.edu/~ivo/RNA/, which predicts secondary structure by using two kinds of dynamic programming algorithms: the minimum free energy algorithm of Zuker and Stiegler (*Nucl. Acid. Res.* 9: 133-148 (1981)) and the partition function algorithm of McCaskill (*Biopolymers* 29, 1105-1119 (1990)). Distances (dissimilarities) between secondary structures can be computed using either string alignment or tree-editing (Shapiro & Zhang 1990). Finally, an algorithm is provided to design sequences with a predefined structure (inverse folding).

Modifications to reduce secondary structure in DNA sequences by altering codon usage can be made in several ways. As used herein, "replacing codons" or "altering codon usage" means altering at least one of the nucleotides making up the three nucleotides of the codon triplet. It is understood that this change can occur at a "wobble" position to leave the amino acid encoded unchanged, or at another position or to a base that results in a change in the encoded amino acid. For example, the codon changes can be designed to swap out codons for a particular amino acid in the sequence (e.g., at a designated position in the sequence) which are not common in the

selected host (following e.g., Kane, *supra*, or Zahn, *supra*). Further, codons can be replaced to reduce the G-C content of the naturally occurring codon.

The inventive methods of the present invention produce sequences with superior expression characteristics because they take more than one variable into account. The methods involve designing a nucleic acid sequence based on a desired amino acid sequence using the codons most commonly used for each amino acid in the chosen host organism (of course, an additional step of analyzing the $\Delta G_{folding}$ of a native sequence may be performed as well). Next, the predicted free energy of folding for the designed sequence is calculated using a computer program as described previously. The program *mfold* is used in the Examples provided herein, although any similar program may be used in the practice of this invention. In calculating the predicted $\Delta G_{folding}$, the full-length nucleotide sequence can be analyzed as a single entity, or the full-length sequence can be divided into shorter segments and the predicted $\Delta G_{folding}$ for each segment can be calculated separately, and then added together.

After the predicted $\Delta G_{folding}$ is calculated, changes to the sequence are made to try to reduce the formation of secondary structure. Regions of predicted secondary structure are identified using, for example, one of the computer programs previously described, and changes are made in codons in these identified regions. Preferably, codon changes are selected to favor more frequently occurring codons in the host organism selected to express the synthetic gene. Thus, one or more codons in regions of predicted high secondary structure are changed to the second or third most commonly used codon choice for the chosen host organism, and the predicted $\Delta G_{folding}$ is recalculated. This process of codon changes and recalculation of the predicted $\Delta G_{folding}$ is repeated until the predicted $\Delta G_{folding}$ of the sequence examined (e.g., the entire sequence or a portion) is increased (made less negative) by greater than about 2%, preferably greater than about 10%, more preferably greater than about 30%, as calculated by $\Delta G_{folding}$/(number of bases in the sequence analyzed). The starting sequence for the step of designing a sequence (e.g., the naturally occurring sequence) is set as 100%. It is likely that the change in $\Delta G_{folding}$ between the starting sequence and the final product will be smaller when the starting sequence is a completely synthetic sequence based solely on preferred codon usage than when the starting sequence is a naturally occurring sequence from a heterologous organism. $\Delta G_{folding}$ for segments analyzed separately can be added to arrive at a $\Delta G_{folding}$ for the entire sequence, or the $\Delta G_{folding}$ for the entire sequence can be determined in a single calculation. Once the $\Delta G_{folding}$ for the entire sequence has been so determined, it is divided by the sequence length in bases to arrive at a uniform measure of $\Delta G_{folding}$ for comparison of sequences of unequal length.

It is also possible that a synthetic sequence may have a more negative $\Delta G_{folding}$ than its counterpart native sequence. This condition may occur when codon choices must be made to accommodate a particular expression host, or when the native sequence has very little secondary structure to begin with. Preferably, this situation occurs in cases where the native sequence does not have a great deal of secondary structure (see, e.g., Example 9 and Fdh2.1 and SynFdh). Regardless, in such cases, the difference in ΔG/base between the native sequence and the more negative synthetic sequence is preferably less than 0.1 kcal/(mol)(base), more preferably less than 0.05 kcal/(mol)(base), and most preferably less than about 0.03 kcal/(mol)(base).

Several variants can be analyzed to illustrate the advantages of the inventive method, summarized in Table 2 below. A naturally occurring (native) *mdeA* gene from *P. putida* (SEQ. ID NO. 1) was used as the starting sequence, and its $\Delta G_{folding}$ was calculated (all $\Delta G_{folding}$ results reported herein were carried out assuming a temperature of 37°C, Na^+ = 1 M, and Mg^{++} = 0) and set at 100%. This sequence was modified by replacing rare arginine codons (termed "*repmdeA*;" modifications derived from Zahn, *supra*) with one found most commonly in *E. coli* (SEQ ID NO. 28). The change in $\Delta G_{folding}$/base from this replacement was 1.9%. A more significant alteration of *mdeA* was performed by replacing all of the rare codons mentioned in Kane, *supra*. This sequence was made by exchanging agg, aga, and cga codons with cgt (arginine), cta codons with ctg (leucine), ata with atc (isoleucine), and ccc with ccg (proline) (termed "*raremdeA*;" SEQ ID NO. 29). As seen in Table 2, below, this exchange also did not significantly impact the ΔG of the sequence, resulting in a change in $\Delta G_{folding}$/base of only 1% as compared to the native sequence. Simply replacing a rare codon does not necessarily increase $\Delta G_{folding}$, and in fact, could lower $\Delta G_{folding}$, creating or failing to resolve problems in transcription or translation, or in amplification by PCR methods.

Because the codons known in the art to be rare and potentially to have an impact on expression did not significantly improve the $\Delta G_{folding}$ of the sequence, all codons of *mdeA*'s open reading frame were exchanged for the most common codons in enteric bacteria from Table 1, above (a sequence termed "*optmdeA;*" SEQ ID NO. 30). The $\Delta G_{folding}$ of this sequence was increased 31.8% by this change compared to *mdeA*, a significant improvement. However, when the sequence *optmde A* was analyzed for regions of predicted secondary structure, replacements of codons in areas of high secondary structure were made to generate the designed sequence *synmdeA* (SEQ ID NO. 3). The predicated $\Delta G_{folding}$ was recalculated for this sequence, and a superior sequence with a greatly improved $\Delta G_{folding}$ was created. In this case, $\Delta G_{folding}$ was increased (made less negative) by 40.7% compared to the starting native sequence. Thus, it is

clear that the inventive methods of developing the synthetic sequences go well beyond any suggestions in the art pertaining to codon exchange.

Table 2

Sequence	ΔG (kcal/mol)	ΔG/base	% Change in ΔG
mdeA (1197 bp)	-256.6	-0.214	0%
repmdeA (1197 bp)	-251.8	-0.210	1.9%
raremdeA (1197 bp)	-254.0	-0.212	1.0%
optmdeA (1200 bp)	-175.5	-0.146	31.8%
synmdeA (1200 bp)	-152.5	-0.127	40.7%

The method described herein of formulating synthetic sequences for improved expression can be used for any nucleic acid sequence, even those being expressed in homologous hosts, or with relatively little predicted secondary structure. Most commonly, however, the need to improve expression will arise when expressing proteins in heterologous hosts. Regardless, any starting sequence, preferably with a $\Delta G_{folding}$ /base of about -0.05 kcal/(mole)(base) or less, and more preferably with a $\Delta G_{folding}$ /base of about -0.15 kcal/(mole)(base) or less, and most preferably with a $\Delta G_{folding}$ /base of -0.2 kcal/(mole)(base) or less can be improved for better expression using the methods of the invention. When a $\Delta G_{folding}$ less than about -0.20 kcal/(mole)(base) or an increase of at least about 2% from the starting sequence is reached, the actual sequence of the synthetic DNA can be physically created. Such physical creation of the designed oligonucleotide sequence can be accomplished by any of the methods known in the prior art, for example by oligonucleotide synthesis, or by the nucleic acid synthesis methods of the invention (described more fully below).

Additionally, the invention takes advantage of the improved secondary structure characteristics of the synthetic nucleic acid for enhanced amplification capability, for example using PCR methods. Some of the same features of native nucleic acid sequence that make them difficult to express in heterologous hosts may also make them difficult to clone or amplify. High secondary structure in one or more regions of the nucleic acid can make cloning or PCR difficult or impossible to perform on the intact nucleic acid or even on segments of the nucleic acid. However, using the methods of the invention to reduce the secondary structure, the resulting nucleic acid templates have better properties for polymerization and amplification. Making

synthetic nucleic acids that amplify easily has important ramifications for common molecular biology procedures such as site directed mutagenesis. For example, using the methods of the invention, a nucleic acid sequence encoding a particular protein (a native protein, a protein with one or more desired mutations, or a completely artificial protein) can be designed using codons used more commonly in a desired expression host cell, and the predicted $\Delta G_{folding}$ may then optimized as described herein. Regardless of the features of the polymerase, or any particular weaknesses it may have (e.g., poor processivity), the probability of accurate full length synthesis of the copy strand from the template is improved using the synthetic nucleic acid of the invention because the regions of secondary structure have been reduced. Codons are replaced overall to minimize $\Delta G_{folding}$ in kcal/(mole)(base), but in specific locations also to alter the amino acid sequence encoded by the nucleotide sequence, resulting in a nucleic acid sequence encoding a particular protein with improved amplification and expression properties.

In one embodiment of this invention, the design and preparation of synthetic genes are used in application of directed evolution, gene shuffling and molecular breeding methods. Examples of gene shuffling and molecular breeding are described in US Patent 5,605,793, US Patent 5,811,238, US Patent 5,830,721, US Patent 5,837,458, US Patent 5,965,408, US Patent 5,958,672, US Patent 6,001,574, all herein incorporated by reference. Genes to be shuffled or recombined are designed and/or physically created based on the incorporation of preferred codons as described in the present invention. Such synthetic genes can also be created with greater homology, improving the reassembly of fragments in gene reassembly and shuffling methods. The advantage of the use of genes designed and physically created as described herein is the improved formation and expression of the shuffled or recombined genes. Such improved expression facilitates screening by providing higher levels of the gene products that are to be detected. The time required for screening can be reduced, or certain enzymatic activities can be detected more easily. Improvements in gene products, whether enzymes or metabolites produced by the actions of two or more different proteins derived through molecular breeding or directed evolution methods, can be detected more readily. Genes designed and produced according to the methods of the present invention can also be incorporated into kits for screening or other purposes. An example of an enzyme screening kit is found in US Patent 6,004,788, herein incorporated by reference.

Another embodiment of the invention, illustrated in the examples below, involves an improved method of synthesizing a nucleic acid. Usual methods of synthesizing a desired nucleic acid sequence which is not found in nature involves difficult and expensive chemical synthesis.

The synthesis method of the invention to create a synthetic sequence involves an amplification method, such as PCR, using synthesized oligonucleotides designed to be overlapping, having as many adjacent sense and antisense strands as desired or required to complete the synthetic gene of choice. The oligonucleotides serve as both the template and primer in this PCR-based synthesis strategy.

The examples described herein demonstrate one implementation of the method for the physical creation of a synthetic gene. Two rounds of PCR reactions were carried out on three segments of the *synmdeA* gene, and six oligonucleotides per segment were used to construct the synthetic gene. The segments were ligated, amplified, excised, and inserted into an expression vector. The first round of amplification involved creating four long oligonucleotides (around 100 bps) based on the synthetic sequence. These long oligonucleotides were used to generate template DNA for various segments of the sequence. Longer synthetic sequences are best broken into shorter segments in this method for easier amplification. The first round PCR amplification relies on overlapping sections of each long oligonucleotide, to create areas of overlap. The areas of overlap serve to prime the extension of the neighboring segment. The areas of overlap can be any length that is sufficient for specificity and long enough for polymerase recognition/attachment, preferably at least 10 bases and more preferably at least 15 bases of overlap.

The second round of amplification used two short oligonucleotides (each about 30 nucleotides) to amplify the full-length segments. The short oligonucleotides overlap the 5' ends of the sense and antisense strands from the previous round to form a template of each segment primed by the first round strands, resulting in the filling in of both 5' and 3' ends after the second round of PCR. The segments derived from this two-round PCR are ligated together to form the unitary synthetic sequence. Preferably, this is facilitated using naturally occurring or synthesized restriction sites. Such sites enhance unidirectional cloning, ligation, etc.

It is understood that any nucleic acid and any reaction conditions that do not require exactly this sort of overlap and/or priming (e.g., RNA, RNA polymerases) can be used to create a modified nucleic acid of the invention without departing from the scope of the invention, and that other means of synthesizing the desired gene of interest are possible using methods known in the art. It is further understood that the gene or nucleic acid can be synthesized in one or several pieces. Likewise, many vectors and host species and strains other than those used herein can be used successfully in the practice of the invention.

The invention is described more fully in the following Examples, which are presented for illustrative purposes only and are not intended to limit the scope of the invention. In the embodiment of the invention disclosed by the Examples, a synthetic gene was designed which encodes the enzyme methionine gamma-lyase. Methods and vectors for its cloning and expression are provided, although other methods/vectors can be used.

EXAMPLE 1 – Design of a synthetic gene sequence

In these Examples, a specific synthetic gene sequence is disclosed encoding naturally occurring *P. putida* methionine gamma-lyase gene sequence, and consists of codons common to enteric bacteria such as *E. coli.* Also described are three gene fragments derived from the complete synthetic methionine gamma-lyase gene that have unique cloning sites at each end of each fragment.

Materials:

DNA taq polymerase and T4 DNA ligase were purchased from Roche (Branchburg, NJ). Restriction endonucleases were purchased from New England Biolabs. Any suitable expression vector, such as pET15b expression vector and *E. coli* BL21(DE3), available from Novagen (Madison, WI), may be used to express the synthetic sequences. pBAD expression vector and *E. coli* LMG 194 were purchased from Invitrogen (Carlsbad, CA). pGEM-3Z, pGEM-5Zf(+) cloning vectors and *E. coli* JM109 were purchased from Promega (Madison, WI). The oligonucleotides for PCR amplification were synthesized by IDT Inc. (Coralville, IA). QIAquick gel extraction kit and QIAprep spin miniprep kit were purchased from QIAGEN, Inc. (Valencia, CA).

Equipment:

Thermocycler Perkin Elmer model 9600 (1991).
Centrifuge
Water bath incubator
Culture incubator
Electrophoresis devices

Software:

mfold - Prediction of RNA secondary structure by free energy minimization; Versions 2.0 and 3.0: suboptimal folding with temperature dependence. Michael Zuker and John Jaeger; Macintosh version developed by Don Gilbert

DNA strider 1.01 - a C program for DNA and protein sequence analysis designed and written by Christian Marck, Service de Biochime-Departement de Biologie, Institut de Recherche Fondamentale Commissariat a I' Energie Atomique- France

HyperPCR - a Hypercard v. 20 stack to determine the optimal annealing temperature for PCR reaction and complementarity between the 3' ends of the two oligos and for internal complementarity of each 3' end. Developed by Brian Osborne, Plant Gene Expression Center, 800 Buchanan St., Albany, CA 94710

Amplify 1.2 - for analyzing PCR experiments. Bill Engels 1992, University of Wisconsin, Genetics, Madison, WI 53706, WREngels@macc.wisc.edu

Lasergene 99 - a complete DNA sequence analysis system. DNASTAR, Inc., 1228 South Park St., Madison, WI 53715.

Design of synthetic DNA sequence encoding Pseudomonas putida methionine gamma-lyase.

The DNA sequence of naturally occurring *mdeA* gene was obtained from Entrez nucleotide Query (NID g2217943) (SEQ ID. NO. 1). Based on this DNA sequence and the amino acid sequence deduced from its open reading frame, several of the original codons were changed to codons that are more commonly used in enteric bacteria. The resulting designed sequence is shown in FIG. 1A (SEQ ID NO. 2). After changing codons to those more commonly used in *E. coli*, the computer program mfold was run to calculate the predicted $\Delta G_{folding}$ the sequence. The computer program was then used to generate an image of the predicted oligonucleotide, and regions of predicted secondary structure were identified. Codons in regions of high secondary structure were changed to the second most commonly used codon for that amino acid in *E coli*, and the predicted $\Delta G_{folding}$ the sequence was recalculated.

In addition, the sequence was modified to incorporate a non-naturally occurring glycine at amino acid position 2. The synthetic sequence therefore does not encode a protein identical to the naturally occurring polypeptide encoded by the *P. putida* methionine gamma-lyase gene. The modification of the sequence was incorporated to facilitate unidirectional cloning of the synthetic sequence into the cloning and expression vectors using an *Nco* I restriction site. The

modified DNA sequence was termed *synmdeA* (SEQ ID NO. 2). In this Example, approximately fifty percent of the codons were changed from those found in the naturally-occurring gene.

EXAMPLE 2 – Amplification of the synthetic DNA fragments *mdeA1, mdeA2, mdeA3*

Oligonucleotide Design:

Oligonucleotide primers were synthesized on the basis of the nucleic sequence of the *synmdeA* gene, whose sequence was determined from the process described in Example 1. The *synmdeA* gene, with 1200 bps of coding sequence (1207 bps with residual bases from restriction sites included) (SEQ ID NO. 3), was broken down into three fragments, *mdeA1*, *mdeA2*, and *mdeA3*. The first cloning fragment, *mdeA1*, contained a *Nco* I cloning site at the 5' end and a *Pst* I cloning site at the 3' end, and was 426 bps after the double stranded product was digested (SEQ ID NO. 4), 441 bps after second round amplification but before digestion (FIG. 1B; SEQ ID NO. 5). The second cloning fragment, *mdeA2*, contained a *Pst* I cloning site at the 5' end and an *Eco*RI cloning site at the 3' end, and was 410 bps after digestion (SEQ ID NO. 6), 430 bps after second round amplification but before digestion (FIG. 1C; SEQ ID NO. 7). The third one, *mdeA3*, contained an *Eco*R I cloning site at the 5' end and a *Bam*H I cloning site at the 3' end, and was 366 bps after digestion (SEQ ID NO. 8), 383 bps after second round amplification but before digestion (FIG. 1D; SEQ ID NO. 9). The segments were the product of internal restriction sites occurring in the *synmdeA* sequence. Restriction sites were chosen that roughly divided the sequence into three equal segments, and which correspond to common multiple cloning sites on commercially available vectors.

To synthesize the segments, or fragments, four long oligonucleotides (98-117 bps), and two short oligonucleotides (~30 bps) were designed for each fragment, and with the help of computer software, their self-folding secondary structures were minimized as much as possible in order to maximize the DNA synthesis during PCR reactions. All the oligonucleotides had secondary structure ΔG's less negative than the ΔG's of the two overlapping annealed fragments, decreasing the probability of secondary structure forming instead of oligonucleotide hybridization.

Two short oligonucleotides and four long oligonucleotides were designed for each of the three segments. They were designed to have 17 to 18 bps overlap with each other. Underlined nucleotides indicate the annealing regions between two adjacent oligonucleotides.

1. First segment of *synmdeA*: *mdeA1*

The sequences of these oligonucleotides was as follows:

mdePr1-1 (33 bps): 5' CAA GAG GC**C** **ATG** GGT CAC GGC TCC AAC AAA CT**G** 3' (sense) (SEQ ID NO. 10)

mdePr1-2 (114 bps): 5' CAC GGC TCC AAC AAA CT**G** CC**G** GG**C** TTT GC**T** ACC CG**C** GCT AT**C** CAC CA**C** GG**T** TA**T** GAC CC**G** CAG GA**T** CAC GG**T** GG**T** GCA CTG GT**T** CC**G** CCG GT**T** TAC CAG AC**T** GC**T** AC**T** TTC ACC 3' (sense) (SEQ ID NO. 11)

mdePr1-3 (116 bps) : 5' GC TTC CAG CAG GTT CAG GGT CGG GTT GGA GAT ACG GGA GTA GAA GTG ACC AGC CTG TTC GCC AGC AAA GCA CGC AGC GCC GTA TTC AAC GGT CGG GAA GGT GAA AGT AGC AGT CTG 3' (antisense) (SEQ ID NO. 12)

mdePr1-4 (117 bps): 5' CT**G** AAC CTG CTG GAA GC**A** CG**T** ATG GC**A** TC**T** CTG GAA GGC GGC GA**A** GC**T** GG**T** CTG GCG CTG GC**A** TC**T** GG**T** ATG GG**C** GCG ATC AC**C** TC**T** AC**C** CT**G** TGG AC**C** CTG CTG CG**T** CC**G** GGT GAC 3' (sense) (SEQ ID NO. 13)

mdePr1-5 (116 bps): 5' GC CAT ATC TAC GTG ACG CAG TTT AAC GCC GAA TTC ACC GAT ACC GTG GTG CAG GAA AGC AAA AGT ACA ACC ATA CAG GGT GTT GCC CAG CAG AAC TTC GTC ACC CGG ACG CAG CAG 3' (antisense) (SEQ ID NO. 14)

mdePr1-6 (33 bps): 5' CAG TGC CTG CAG GTC AGC CAT ATC TAC GTG ACG 3' (antisense) (SEQ ID NO. 15)

2. Second segment, *mdeA2*

The sequences of these oligonucleotides was as follows:

mdePr2-1 (33 bps): 5 ' GCT GAC CTG CAG GCA CTG GAA GCG GCT ATG ACC 3' (sense) (SEQ ID NO. 16)

mdePr2-2 (114 bps): 5' CTG GAG GCT GCT ATG ACC CCG GCT ACC CGT GTT ATC TAC TTC GAA TCC CCG GCT AAC CCG AAC ATG CAC ATG GCT GAC ATC GCA GGT GTT GCT AAA ATC GCT CGT AAG CAC GGC 3' (sense) (SEQ ID NO. 17)

mdePr 2-3 (115 bps): 5' G GTA TTT AGT AGC GGA GTG AAC AAC CAG GTC AGC GCC CAG TTC CAG CGG ACG TTG CAG GTA CGG AGT ACA GTA GGT GTT ATC AAC AAC TAC GGT AGC GCC GTG CTT ACG AGC GAT 3' (antisense) (SEQ ID NO. 18)

mdePr2-4 (111 bps): 5' CAC TCC GCT ACT AAA TAC CTG TCC GGC CAC GGC GAC ATC ACT GCT GGC ATC GTA GTA GGC TCC CAG GCA CTG GTT GAC CGT ATC CGT CTG CAA GGT CTG AAA GAC ATG ACC 3' (sense) (SEQ ID NO. 19)

mdePr2-5 (115): 5' G TAC CTG AGC GTT AGC ACA GTG ACG GTC CAT ACG CAG GTT CAG GGT CTT GAT ACC ACG CAT CAG CAG TGC TGC GTC GTG CGG GGA CAG AAC AGC GCC GGT CAT GTC TTT CAG ACC 3' (antisense) (SEQ ID NO. 20)

mdePr2-6 (33): 5' C CAG GAA TTC AGC CAG TAC CTG AGC GTT AGC AC 3' (antisense) (SEQ ID NO. 21)

3. Third segment, *mdeA3*

The sequences of these oligonucleotides was as follows:

mdePr3-1 (31 bps): 5' T CTT AAT GAA TTC CTG GCT CGT CAG CCG CAG 3' (sense) (SEQ ID NO. 22)

mdePr3-2 (105 bps): 5' CTG GCT CGT CAG CCG CAG GTA GAA CTG ATC CAC TAT CCG GGC CTG GCT TCC TTC CCG CAG TAC ACT CTG GCA CGT CAG CAG ATG TCC CAG CCG GGC GGT ATG ATC 3' (sense) (SEQ ID NO. 23)

mdePr3-3 (106 bps): 5' C GTC ACC CAG GGA AAC CGC ACG GGA GAA CAG CTG CAG AGC GTT CAT GAA ACG ACG ACC AGC GCC GAT GCC ACC CTT CAG TTC GAA AGC GAT CAT GCC ACC CGG CTG 3' (antisense) (SEQ ID NO. 24)

mdePr3-4 (106 bps) 5' GCG GTT TCC CTG GGT GAC GCT GAA TCC CTG GCG CAG CAC CCG GCA TCC ATG ACT CAC TCC TCC TAC ACT CCG GAA GAA CGT GCG CAC TAC GGC ATC TCC GAA GGC C 3' (sense) (SEQ ID NO. 25)

mdePr3-5 (98 bps): 5' CA AGC GCT AGC CTT CAG AGC CTG CTG AAC GTC TGC CAG CAG ATC ATC GAT GTC TTC CAG ACC AAC AGA CAG ACG AAC CAG GCC TTC GGA GAT GCC GTA 3' (antisense) (SEQ ID NO. 26)

mdePr3-6 (32 bps): 5' T GGT GGA TCC TCA AGC GCT AGC CTT CAG AGC C 3' (antisense) (SEQ ID NO. 27)

Amplification of segmental DNA: *mdeA1*, *mdeA2*, *mdeA3*:

Each segment synthesis took two rounds of amplification. The first round was to generate the template for the second round using the four long oligonucleotides with overlapping ends (e.g., 3' or 5' sense ends overlapping neighboring 5' or 3' antisense ends). The second round amplification was using the two short nucleotides and the template from the first. Standard PCR reaction mixture was used with 100 µl reaction volume, 0.2 mM dNTPs (final concentration), and 60 to 90 pmoles of each oligonucleotide.

To synthesize the template for *mdeA1*, termed *tpA1*, mdePr1-2 (71 pmoles), mdePr1-3 (74 pmoles), mdePr1-4 (77 pmoles), and mdePr1-5 (64 pmoles) were used. MdePr2-2 (64 pmoles), mdePr2-3 (73 pmoles), mdePr2-4 (67 pmoles), and mdePr2-5 (74 pmoles) were used to synthesize *mdeA2* template, termed *tpA2*. To synthesize *mdeA3* template, termed *tpA3*, mdePr3-2 (66 pmoles), mdePr3-3 (62.6 pmoles), mdePr3-4 (60 pmoles), and mdePr3-5 (82 pmoles) were used. The strategy is shown in FIG. 2A. Based on the estimated annealing temperatures between the oligonucleotides above, the PCR reaction conditions were as follows:

first denaturation at 94°C for 2 min; then 10 cycles of denaturation at 94°C for 30 sec; annealing at 51°C for 40 sec, and extension at 72°C for 1 min. This was followed by 20 cycles of denaturation at 94°C for 30 sec; 65°C for 55 sec; 72°C for 1 min; then a final extension at 72°C for 7 min. The PCR was carried out using a Perkin-Elmer Gene Amp 9600.

The PCR products were separated on 2 % agarose gels run with a 1 kb DNA ladder (NEB); product bands of the expected size (411 bps for *tpA1,* 401 bps for *tpA2,* and 360 bps for *tpA3*) were cut out and extracted using QIAquick gel extraction kit. The products were then used as the templates for second round PCR reactions to synthesize *mdeA1, mdeA2,* and *mdeA3* DNAs. The strategy for the second round amplification is shown in FIG. 2B.

For the second round, mdePr1-1(80 pmoles), mdePr1-6 (67 pmoles), and 1 µl of 50 µl gel purified template *tpA1* (above) were used to amplify the *mdeA1* segment, again with the 3' end of mdePr1-1 and mdePr1-6 overlapping the 5' end of the template, and each 3' end (of oligonucleotide or template) priming the extension of the full length segment product. Similarly, mdePr2-1 (86 pmoles), mdePr2-6 (86 pmoles), and 1 µl template *tpA2*; mdePr3-1 (74 pmoles), mdePr3-6 (84 pmoles), and 1µl *tpA3* were used to amplify *mdeA2* and *mdeA3* segment respectively. The PCR reaction conditions were as follows: first denaturation at 94°C for 2 min; then 25 cycles of denaturation at 94°C for 30 sec, annealing at 51°C for 40 sec, and extension at 72°C for 30 sec; followed by a final extension at 72°C for 7 min.

The PCR-amplified products were identified by size on the 2 % agarose gel, a 441 bp-band for *mdeA1,* a 430 bp-band for *mdeA2*, and a 383 bp-band for *mdeA3.* The DNAs from the bands were extracted by using QIAquick gel extraction kit.

EXAMPLE 3 – Cloning the synthetic DNA fragments *mdeA1, mdeA2,* and *mdeA3* into an appropriate vector

The vector pGEM-5Z (Promega, 3003 bps), and the purified PCR *mdeA1* DNA were double cut with *Nco* I and *Pst* I; pGEM-3Z (Promega, 2743 bps), and the purified PCR *mdeA2* DNA were double cut with *Pst* I and *Eco*R I restriction enzymes; pGEM–3Z and purified PCR *mdeA3* DNA were double cut with *Eco*R I and *Bam*H I restriction enzymes. These vectors carry the multiple cloning site arrangement from pUC18, and are ampicillin resistant. All restriction digestion reactions were incubated overnight at 37°C. The digested products were then purified by gel electrophoresis on a 2% agarose gel followed by extraction of the DNA using a QIAquick gel extraction kit.

The purified, double cut pGEM-5Z and *mdeA1* were ligated with T4 DNA ligase and buffers (NEB) and incubated overnight at 16°C. Similarly, the double cut pGEM-3z and *mdeA2*, and double cut pGEM-3z and *mdeA3*, were ligated with T4 DNA ligase, but they were incubated at 12°C because *Eco*R I site requires lower temperature to anneal. Several reactions were carried out for each construct to ensure optimization of molar ratios between vector and insert (e.g. 1:1, 1:3, and 3:1 vector : insert ratio). FIG. 3 illustrates the multiple cloning site and ligation of inserts into the vectors.

E. coli JM109 competent cells (Promega or Bio 101) were transformed with the ligation reactions described above using a standard heat shock transformation procedure (Sambrook et al., 1989, *supra*). To select for colonies containing *mdeA1*, *mdeA2*, and *mdeA3* clones, the cells were grown on LB+Ampicillin (50μg/μl) plates.

Transformant colonies were first tested with PCR screening using the mdePr1-1, mdePr1-6, mdePr2-1, mdePr2-6, mdePr3-1, and mdePr3-6 as the primers for *mdeA1*, *mdeA2*, and *mdeA3* clones respectively. The PCR reaction volume was 25 μl with 0.2 mM dNTPs and 20 pmoles of each primers. The templates were picked directly from the colonies, and the conditions were as follows: first denaturation at 94°C for 4 min; then 25 cycles of denaturation at 94°C for 30 s; annealing at 57°C for 40 s; and extension at 72°C for 30 s; then a final extension at 72°C for 7 min. The positive colonies containing *mdeA1*, *mdeA2*, or *mdeA3* clones were identified by the presence of 441 bp, 430 bp, or 383 bp bands respectively.

To further confirm that the colony actually carried the *mdeA1*, *mdeA2*, or *mdeA3* construct, restriction mapping of its plasmid was done by cutting the plasmid with *Nco* I + *Pst* I, *Pst* I + *Eco*R I, or *Eco*R I + *Bam*H I. The presence of a 426 bp-band (*mdeA1*), a 414 bp-band (*mdeA2*), or a 367 bp-band (*mdeA3*) would be expected on 2 % agarose gel if the plasmid carries the proper insert.

EXAMPLE 4 – Sequencing of the synthetic *mdeA1*, *mdeA2*, and *mdeA3* DNA fragments

After isolating plasmids containing the *mdeA1*, *mdeA2* and *mdeA3* inserts, the clones were submitted to the UCLA sequencing facility (Los Angeles, CA) for sequencing. M13 forward and reverse primers were used. Clones that carried the correct DNA sequence of *mdeA1*, *mdeA2*, and *mdeA3* were selected and named *pSmA1-17*, *pSmA2-8*, and *pSmA3-3*.

EXAMPLE 5 – Construction of full-length *synmdeA* encoding methionine gamma-lyase

The colonies containing *pSmA1-17, pSmA2-8,* and *pSmA3-3* were cultured with LB+ampicillin (50μg/μl) overnight at 37 C. Plasmids were extracted using QIAprep spin miniprep kit (QIAGEN, Inc., Valencia, CA). The plasmids *pSmA1-17, pSmA2-8,* and *pSmA3-3* were double cut overnight at 37°C with *Nco* I/*Pst* I, *Pst* I/*Eco*R I, and *Eco*R I/ *Bam*H I restriction enzymes respectively. A pET15b vector (Novagen) was cut with *Nco* I/ *Bam*H I restriction enzymes, and a pBAD/His C vector (Invitrogen) was cut with *Nco* I/Bgl II. The double cut DNAs were separated on 2% agarose gel, and the bands corresponding to *mdeA1* (426 bps), *mdeA2* (414 bps), *mdeA3* (367 bps), pET15b (5k bps), and pBAD/His C (4 kbs) were isolated and purified using QIAquick gel extraction kit.

Purified *mdeA1*, *mdeA2*, and *mdeA3* DNAs were then ligated into double cut pET15b at *Nco* I and *Bam*H I, or pBAD/His C at *Nco* I and Bgl II cloning sites using T4 DNA ligase overnight at 12°C.

The resulting plasmids were transformed into E. coli JM109 competent cells using a standard heat shock transformation procedure (Sambrook et al., 1989, *supra*). To select the positive clones containing *synmdeA*, the cells were grown on LB+Ampicillin (50μg/μl) plates overnight at 37°C.

The transformant colonies were first checked with the PCR screening method described above by using mdePr1-1 and mdePr3-6 as the primer probes. A 1200 bp-band was expected on the agarose gel if the colony contained *synmdeA* clones. Selected pET15b and pBAD/His C vectors carrying the *synmdeA* insert were named pTM-1 and pBM-1 overexpression plasmids, respectively. The PCR positive colonies were then further confirmed by using a restriction mapping method, with *Nco* I and *Bam*H I restriction enzymes used on pTM-1, and *Nco* I and *Hind* III restriction enzymes used on pBM-1. Again, 1200 bp-bands were seen on 2% agarose gels.

Plasmids pTM-1 and pBM-1 were transferred to expression host *E. coli* BL21(DE3) and LMG 194 by first plasmid extraction, followed by transformation.

EXAMPLE 6 – Over-expression of synthetic L-Methionine-alpha-gamma-lyase gene

Host *E. coli* strains carrying pTM-1 and pBM-1, referred to as BL/pTM01 and LMG/pBM01 respectively, were grown on LB+ampicillin plate and RMG+ampicillin plate respectively. A single colony from each plate was then picked and cultured overnight in LB+ampicillin liquid medium. Then 5 ml of LB+ampicillin was inoculated with 100 µl of each overnight culture, and each was incubated for 2 hours at 37°C with shaking or until $O.D._{600}$ (nm) reached 0.8 – 0.9. Initially, 1 ml of each culture was removed as a non-induced control. BL/pTM01 culture was then induced to express protein by adding IPTG to a final concentration of 2 mM, and LMG/pBM01 culture was induced with a final concentration of 0.02% L-arabinose. Incubation was continued at 37°C for 3 hours. Samples of 1 ml were collected every hour. All samples were centrifuged at 12,000 x g for 3 minutes. The cells were then lysed by resuspension in 1x NuPAGE sample buffer (Novex) containing 50 mM DTT, and incubation at 97°C for 3 minutes. After centrifugation for 10 min at 12,000 x g, the supernatants were separated along with protein size markers by SDS-page on 4%-20% gradient polyacrylamide gel (NuPAGE MES SDS, Novex) for 1 hour at 150 volts. The gels were stained by Coomassie blue for 2 hours and destained in 10% acetic acid, 20% methanol solution, followed by destaining in 7% acetic acid, 5% methanol. 43 kD bands corresponding to a molecular weight marker were seen on the destained gels (FIG. 4). These bands corresponded to the major protein in the induced samples. As seen in FIG. 4, expression of *synmdeA* was vastly superior to expression of the native enzyme, seen in FIG. 5. The native enzyme expressed poorly in *E. coli*, and was a truncated portion of the complete gene. Attempted expression of the native gene gave a protein of apparent molecular weight approximately 28kD, indicating that a substantial part of the enzyme was missing. The protein showed no methionine gamma-lyase activity. Without wishing to be bound to any particular mechanism, it is hypothesized that the truncation was caused by an interruption in translation at a rare codon. This speculation is supported by the fact that an interruption at this point would result in a polypeptide product having a molecular weight of approximately 28 kD.

EXAMPLE 7 - Comparison of native *mdeA* and *synmdeA* gene expression

To demonstrate the usefulness of the synthetic gene for the expression of difficult to express genes in *E. coli*, the *synmdeA* gene was expressed in *E coli* using the vector pET15b. This gene encodes a methionine β lyase enzyme, but contains an additional amino acid relative to the native protein described by Soda and co-workers (e.g., US Patent No. 5,863,788). The results are shown in the gel in FIG. 4A. Based on the density of the band corresponding to the

methionine-gamma lyase enzyme of approximate molecular weight 40,000 we estimate the level of expression to be 10% or more of the total protein in the crude cell lysate of the *E. coli* host. By contrast, expression of the native *mdeA* gene in the vector pSIT is substantially less under the same induction conditions (FIG. 4B). In the experiment shown in FIG. 4B, all samples were incubated at 37°C. The induced samples contain extra bands of about 28 kD which indicate that premature termination of the enzyme occurred during translation of the native gene. Both the native and synthetic gene vectors are under the control of T7 RNA polymerase promotors.

To put these results into another context, the expression reported by Soda and coworkers in US Patent 5,861,154 and 5,863,788 is reported to be 0.82 units/mg. Using the specific activity of the purified enzyme of 20.4 units/mg reported by Soda in *Anal. Biochime.* 138, 421-424 (1984), the expression level is estimated to be no more than 4% of the total protein in the *E. coli* host. This estimate is an upper limit on the expression reported by Soda because the reported activity involves some partial purification of the enzyme prior to assay.

EXAMPLE 8 – Comparison of expression of genes with different $\Delta G_{folding}$

FIG. 5 is a gel showing expression of two genes with different $\Delta G_{folding}$. Naphthalene Dioxygenase from *P. putida* has a $\Delta G_{folding}$ of -256.1 kcal/mol. This very low free energy would not be expected, under the principles of the invention, to express well. In fact, as seen in lanes 1-4 of FIG. 5, it does not. By contrast, another gene, methionine gamma lyase (*mgl I*) from *T. vaginalis* has a $\Delta G_{folding}$ of -152.5 kcal/mol. As can be seen from lanes 6-9 of FIG. 5, this protein can be induced and expresses well under the conditions used. Both genes were cloned into the pBAD vector and grown at 37°C.

EXAMPLE 9 – Synthesis of improved eukaryotic genes and their expression in prokaryotic hosts

Oxidoreductases

The enzyme family of oxidoreductases is large and complex, and many members function to stereoselectively oxidize and reduce functional groups such as C=O, C=C, and C=N. In pharmaceutical and agricultural industries, for example, these enzymes are used to prepare drugs and chemicals requiring e.g., chiral compounds. For example, they can be used to stereoselectively reduce ketones to produce chiral alcohols consisting predominantly of a single stereoisomer. In this Example, the methods of the invention were used to create highly expressible oxidoreductases. Properties of exemplary original oxidoreductase genes and their

synthetic analogs are discussed and shown in Table 3 below, and the superiority of the synthetic sequences in ΔG, expression, and enzyme activity can be seen.

Keto Reductases

These enzymes reduce keto esters, aldehydes, and other ketones into equivalent alcohol products.

NADPH-Dependent Aldehyde Reductase 1, ALR1: The native gene encoding a NADPH-dependent aldehyde reductase (ALR) is from a red yeast, *Sporobolomyces salmonicolor* (also known as *Sporidiobolus salmonicolor)*, and catalyzes the reduction of a variety of carbonyl compounds. The gene is 969 bp (SEQ ID NO. 31) and encodes a polypeptide of 35,232 Da. The deduced amino acid sequence (SEQ ID NO. 32) shows a high degree of similarity to other members of the aldo-keto reductase superfamily. The synthetic aldehyde reductase 1 gene (synALR1; SEQ ID No. 33) was created using the known protein sequence.

Aldehyde Reductase 2, ALR2: This gene, encoding an NADPH-dependent aldehyde reductase (AR2) in *Sporobolomyces salmonicolor* AKU4429, reduces ethyl 4-chloro-3-oxobutanoate (4-COBE) to ethyl (S)-4-chloro-3-hydroxybutanoate (Kita *et al., Appl Environ Microbiol* 1999 Dec; **65(12):**5207-11). The ALR2 gene (SEQ ID NO. 34) is 1,032 bp long and encodes a 37,315-Da polypeptide. The deduced amino acid sequence (SEQ ID NO. 35) exhibits significant levels of similarity to the amino acid sequences of members of the mammalian 3-beta-hydroxysteroid dehydrogenase-plant dihydroflavonol 4-reductase superfamily but not to the amino acid sequences of members of the aldo-keto reductase superfamily or to the amino acid sequence of an aldehyde reductase previously isolated from the same organism (*K. Kita, et al., Appl. Environ. Microbiol.* **62**:2303-2310, 1996; SEQ ID NO. 32). The synthetic version of ALR2, or synALR2mut (SEQ ID NO. 36) contains a mutation at position 25 of the amino acid sequence (SEQ ID NO. 37), replacing alanine with glycine to introduce a mutation that allows the enzyme to use both NADH and NADPH as a cofactor.

Reductase 1 from yeast, YPR1: This enzyme is a good general ketone reductase. The "native" sequence, related to Accession No. X80642 (Miosga *et al.*), was cloned into pBAD with a GGT insertion after the initiating ATG (SEQ ID NO. 38). This addition resulted in a glycine at position 2 in the amino acid sequence in both the "native" and the synthetic YPR1 peptide sequence (SEQ ID NO. 39) to add a restriction site for ease of cloning. SEQ ID NO. 40 is the synthetic sequence, having a 15.1% improvement in $\Delta G_{folding}$.

Yeast GCY1: SEQ ID NO. 41 is a nuclear gene for a yeast protein showing unexpectedly high homology with mammalian aldo/keto reductases as well as with p-crystallin, one of the prominent proteins of the frog eye lens. The coding region is 939 bases and encodes a protein of 312 amino acids (SEQ ID NO. 42; estimated MW 35,000). A synthetic analog was made, synGCY1 (SEQ ID NO. 43), having a GGC insertion after ATG (to facilitate cloning into the pBAD vector), which results in the insertion of a glycine after the initiating methionine in the synthetic peptide sequence (SEQ ID NO. 44).

Reductase Gre2 from yeast: This gene and related protein product were originally sequenced as part of the yeast genome (Goffeau *et al.*, Accession Nos. NC_001147 and NP_014490). The native gene (SEQ ID NO. 45) was not cloned, and its protein sequence (SEQ ID NO. 46) is based on the best open reading frame. However, the synthetic gene synGRE2 (SEQ ID NO. 47) derived from the wild-type sequence was modified by addition of a GGC insertion (to add a restriction site), cloned, and expressed as a protein (SEQ ID NO. 48). The protein's reductase function has been confirmed.

Yeast Aldo-Keto Reductase Gre3: This gene and related protein encode a keto-aldose reductase (Goffeau *et al.*, Accession Nos. NC_001140 and NP_011972). The "native" sequence (SEQ ID NO. 49) has been modified to insert an ATT at the second codon position (inserting isoleucine in SEQ ID NO. 50) to add a restriction site for cloning. The "native" Gre3 protein exhibits reductase activity at 30°C and 37°C as shown in Table 3 below.

CMKR (S1): The product of this gene (SEQ ID NO.69) is an NADPH-dependent carbonyl reductase (S1) from *Candida magnoliae,* which catalyzes the reduction of ethyl 4-chloro-3-oxobutanoate (COBE) to ethyl (S)-4-chloro-3-hydroxybutanoate (CHBE), with a 100% enantiomeric excess. This is a useful chiral building block for the synthesis of pharmaceuticals. The S1 gene is 849 bp and encodes a polypeptide of 30,420 Da. The deduced amino acid sequence (SEQ ID NO. 70) has a high degree of similarity to those of other members of the short-chain alcohol dehydrogenase superfamily.

Table 3: Properties of Native and Synthetic Genes

Gene name	length (bps)	Molecular Weight (kD)	ΔG (kcal/mole)	ΔG/base (kcal/mole·base)	% ΔG difference between native and synthetic	Activity at 30°C (u/ml)	Activity at 37°C (u/ml)
nativeALR1	972	35.2	-152.5	-0.157	100	ND	ND
synALR1	972	35.2	-85.8	-0.0883	56.3	75.5	9.25
nativeALR2	1032	37.3	-162.2	-0.1572	100	ND	ND
synALR2mut	1032	37.3	-101.2	-0.0981	62.4	4.13	7.0
nativeYPR1	942	34.8	-89.4	-0.0949	100	4.15	6.23
synYPR1	942	34.8	-75.9	-0.0806	84.9	11.791	16.609
nativeGCY1	939	35.1	-76.6	-0.0816	100	0.105	0.533
synGCY1	942	35.1	-73.2	-0.0777	95.2	4.00	4.53
nativeGRe2	1029	38.2	-103.3	-0.1004	100	ND	ND
synGRE2	1032	38.2	-71.6	-0.0694	69.1	ND	ND
nativeGRE3	987	37.2	-89	-0.0902	100	0.35	0.52
synGRE3	987	37.2	-65.5	-0.0664	73.6	1.2	1.1
native CMKR	852	30.6	-145.4	-0.1706	100	ND	ND
synCMKR	852	30.6	-70.5	-0.0827	48.5	ND	239.16
pKDDC	1461	54.0	-244.4	-0.1673	100	ND	ND
synAAAD	1464	54.0	-133.9	-0.0915	54.7	ND	ND
Fdh1.2	1098	40.6	-76.1	-0.0693	100	0.48	0.54
synFdh	1098	40.6	-98	-0.0893	128.9	2.48	0.19

ND = not determined.

Other Sequences

L-Aromatic Amino Acid Decarboxylase from Pig Kidney: L-Aromatic amino acid decarboxylase (dopa decarboxylase; DDC) is a pyridoxal 5'-phosphate (PLP)-dependent homodimeric enzyme that catalyzes the decarboxylation of L-dopa and other L-aromatic amino acids. A cDNA that codes for the protein from pig kidney was cloned by Moore *et al., Biochem J* 1996 Apr 1;315 (Pt 1):249-56. Using this pKDDC sequence (SEQ ID NO. 53; Accession No. S82290) and its deduced amino acid sequence (SEQ ID NO 54), a synthetic decarboxylase, synAAAD was constructed with a GGT insertion (SEQ ID NO 55) to insert a glycine in the amino acid sequence (SEQ ID NO. 56). The synAAAD nucleic acid sequence had a nearly 50% improvement in ΔG (see Table 3).

Formate Dehydrogenase (Fdh1.2): The formate dehydrogenase (Fdh1.2) DNA (SEQ ID NO. 57) and protein sequence (SEQ ID NO. 58) is from *Candida boidinii* (Accession No. AJ245934). In order to create a *Nco* I restriction site for cloning into expression vector

pBAD/HisA, a glycine codon was inserted after the first methionine codon (SEQ ID NO. 59). The resultant recombinant protein, synFdh (SEQ ID NO. 60) has an inserted glycine after the initiating methionine as compared to the native protein. Native Fdh1.2 and synFdh otherwise encode the same protein sequence. The synthetic sequence had 199 out of 366 codons changed as compared to native Fdh1.2 to optimize expression in *E. coli* (see Table 4 below). Homology at the DNA level of Fdh1.2 and synFdh is about 78.5%. Expression of synFdh is 5-fold higher based on activity measurements than expressed native Fdh1.2.

The ΔG of Fdh1.2 is -76.1 kcal/mole (-0.069 kcal/mol·base) and the ΔG of synFdh is -98.0 kcal/mole (-0.089 kcal/mol·base). Because native Fdh1.2 does not have high secondary structure, it was possible to optimize the sequence for expression according to methods of the invention without increasing, and in fact, slightly decreasing, the $\Delta G_{folding}$. FIG. 7 shows expression data of Fdh1.2 compared with synFdh at 30°C and 37°C. As shown in FIG. 8, synFdh, induced with 0.2% L-arabinose at 30°C, exhibits higher catalytic activity than does induced native Fdh1.2 or uninduced Fdh1.2 in the oxidation of formate in the presence of NAD^+ ($NAD^+ + HCO_2^- \rightarrow NADH + CO_2$). These figures demonstrate the superior expression characteristics of the synthetic Fdh sequence as compared to the native sequence.

Table 4: Codon Preference of *C. boidinii* and *E. coli* for Selected Amino Acids

Amino Acid	*C. boidinii* Codon	*E. coli* Codon
R (Arg)	AGA (13/13)	AGA (0); CGT (0.74)
N (Asn)	AAT (14/16)	AAT (0.06), AAC (0.94)
D (Asp)	GAT (22/24)	GAT (0.33), GAC (0.67)
Q (Gln)	CAA (9/9)	CAA (0.14), CAG (0.86)
L (Leu)	TTA (22/32)	TTA (0.02), CTG (0.83)
P (Pro)	CCA (11/14)	CCA (0.15), CCG (0.77)
T (Thr)	ACT (11/22)	ACT (0.35), ACC (0.55)

Hydantoinase: The hydantoinase gene from *Pseudomonas putida* (SEQ ID NO. 61) and its deduced amino acid sequence (SEQ ID NO. 62) (Accession No. AAC00209) were used to create synthetic hydantoinase gene (SEQ ID NO. 63) and protein (SEQ ID NO. 64) products. This gene product is useful to make non-natural α-amino acids. To create the synthetic gene and protein, a glycine was added after the first methionine so that the gene could be subcloned into

the pBAD/HisA expression vector. The nucleic acid sequence is 1491 bp, and in its native form has a free energy of folding of -287.6 kcal/mole. The synthetic hydantoinase has a ΔG of -155.5 kcal/mole. Homology at the nucleic acid level between the native and synthetic hydantoinase is 78.4%.

Vanillyl Alcohol Oxidase, VaoA: A vanillyl-alcohol oxidase gene (SEQ ID NO. 65) and its deduced amino acid sequence (SEQ ID NO. 66) from *Penicillium simplicissimum* was used. VaoA oxidizes vanillyl alcohol and related aromatic alcohols. To create the synthetic gene (SEQ ID NO. 67) and protein (SEQ ID NO. 68), a glycine was added after the first methionine so that the gene could be subcloned into the pBAD/HisA expression vector. The sequence is 1686 bp long and the native form has a ΔG of -176.8 kcal/mole; ΔG of synVaoA is -164.6 kcal/mole. The genes have 77% homology at the nucleic acid level.

Myo-Inositol-1-Phosphate Synthase (Ino1): INO-1 (SEQ. ID NO. 73) cyclizes D-glucose 6-phosphate to myo-inositol 1- phosphate, which is a precursor for coenzyme Q. The native *ino-1* gene (SEQ. ID NO. 72) is 1602 bps. The ΔG is -152.2 kcal/mole. The synthetic *ino-1* gene, called *synIno-1* (SEQ. ID NO. 74) has a GGT insertion to create the cloning site, which inserts a glycine residue in the synINO protein (SEQ. ID NO. 75). *synIno-1* is 1605 bps and has a ΔG of –131.8 kcal/mole. The similarity at DNA level of the *ino 1* and *synIno 1* is 77.4 %.

Galactose Oxidase (GAO): The gaoA gene (SEQ. ID. NO. 76), encoding the secreted copper-containing enzyme galactose oxidase (SEQ. ID. NO. 77), was isolated from the Deuteromycete fungus *Dactylium dendroides* (Accession number: M86819; also called *Hypomyces rosellus*). ΔG for the native DNA is –244 kcal/mole and ΔG for the synthetic gene (*synGAO,* SEQ.ID. NO. 78) is – 195.3 kcal/mole. The open reading frame for galactose oxidase (GAO) is 2046 bp. At the DNA level, synGAO and GAO have 76.6% identity. Glactose oxidase oxidizes galactose, and can be used in the quantitative determination of galactose level in blood. The synthetic galactose oxidase protein (SEQ. ID NO. 79) has a glycine inserted in the second amino acid position.

The Gibbs free energy (ΔG) of all DNA foldings described in this Example were determined using *mfold2* provided by Washington University School of Medicine (http://mfold2.wustl.edu). The conditions used for calculation of the free energy of DNA folding were 37°C , Na^{+}= 1M and Mg^{++}= 0.

Assays of enzyme activities of the keto reductases were determined photometrically using Ethyl 4-chloroacetoacetate as a substrate. The reaction mixture (1.0 ml) comprised 50 mM potassium phosphate buffer (pH 6.5), 250 M NADPH, 5mM substrate, and cell lysate. The

reaction was measured at room temperature. One unit of the enzyme was defined as the amount catalyzing the oxidation of 1 mole NADPH/min. Formate dehydrogenase activity was assayed by mixing sodium formate with NAD+, and measuring NADH recycling activity on a spectrophotometer at 340 nm.

As seen by the results generated in this example, the methods of the invention are widely applicable to unrelated genes from both prokaryotes and eukaryotes, and result in improved expression and enzymatic activity when expressed in a heterologous prokaryotic or eukaryotic host cell.

The preceding description has been presented with references to presently preferred embodiments of the invention. Persons skilled in the art and technology to which this invention pertains will appreciate that alterations and changes in the described genes, proteins, and methods can be practiced without meaningfully departing from the principle, spirit and scope of this invention.

Accordingly, the foregoing description should not be read as pertaining only to the precise genes, proteins, and methods described and shown in the accompanying drawings, but rather should be read as consistent with and as support for the following claims, which are to have their fullest and fairest scope.